COMMENT MONITORER LES PRATIQUES DE PARTAGE DES DONNÉES DE LA RECHERCHE ?

PREMIÈRE TENTATIVE DU CÔTÉ DE LA HES-SO

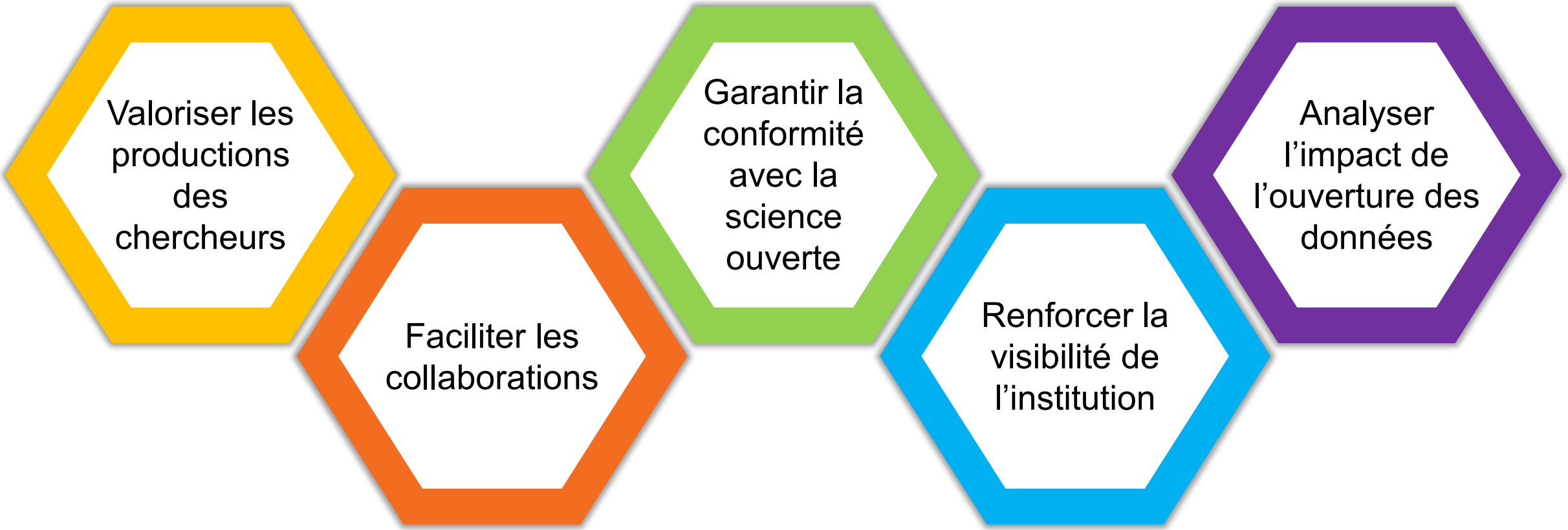
WEBINAIRE LOVE DATA WEEK 2025

10 FÉVRIER 2025

Emilie Pasche

INTRODUCTION

Pourquoi monitorer les pratiques de partage des données de la recherche ?



Monitoring des pratiques du partage des données de la recherche: le cas de la HES-SO

Objectif

Identifier et cartographier les jeux de données et codes/logiciels déposés par les chercheur·euses de notre institution dans des dépôts ouverts pour mesurer les pratiques de partage au sein de la HES-SO

Problème

Contrairement aux publications, il n'y a pas encore de dispositif qui rassemble et valorise les jeux de données et codes/logiciels produits au sein des 28 hautes écoles de la HES-SO

Hes·so

Pourquoi est-il difficile de monitorer les pratiques de partage des données de la recherche sans dépôt institutionnel?

Multiplicité des dépôts

Le FNS a identifié plus de 146 dépôts différents dans les DMPs de 1500 demandes de fonds*: dépôts institutionnels, dépôts généralistes ou dépôts spécifiques.

zenodo
ENA European Nucleotide Archive
HARVARD Dataverse
DaSCH
DRYAD
YARETA
SWISS base

Services d'agrégation

Ces services permettent de centraliser des informations issues de multiples dépôts, offrant ainsi une vue globale et interopérable des productions.

DataCite Commons
OpenAlex
OpenAIRE

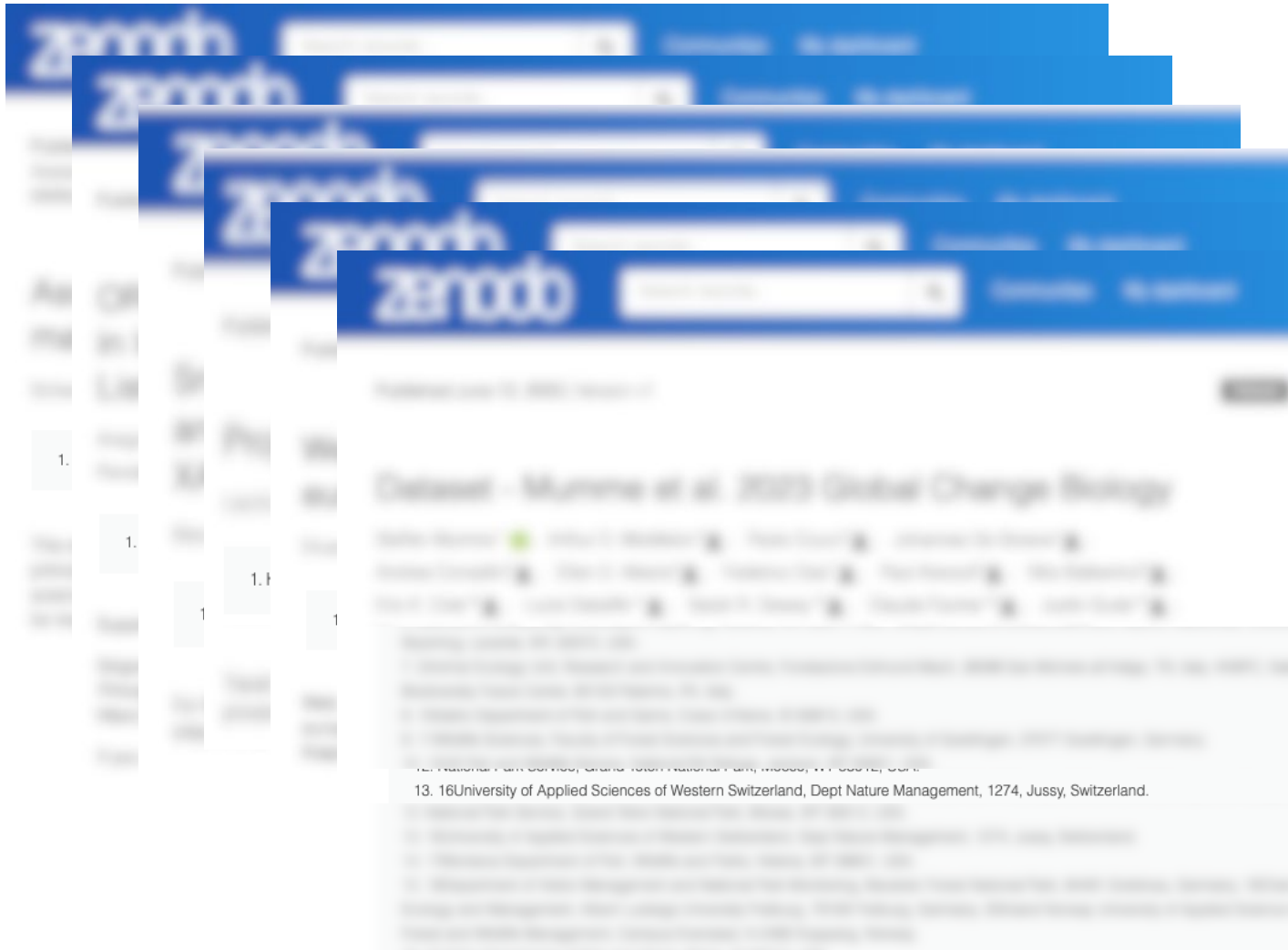
*https://www.snf.ch/en/WtezJ6qxuTRnSYgF/topic/open-research-data-which-data-repositories-can-be-used. Février 2020

Hes·so

Pourquoi est-il difficile de monitorer les pratiques de partage des données de la recherche sans dépôt institutionnel?

Manque de standardisation

Les métadonnées des jeux de données ou codes/logiciels ne sont pas toujours homogènes ou conformes à des standards, notamment concernant les affiliations.
Ce problème est particulièrement complexe pour une institution comme la HES-SO, composée de 28 hautes écoles: nom de l'institution, noms des écoles, noms des instituts, langue des affiliations, formes abrégées ou étendues, etc.



Notre approche

SOLUTION

Exploitation des outils et services d'agrégation, tels que DataCite et OpenAlex, pour collecter les jeux de données et codes/logiciels déposés dans des dépôts ouverts par les chercheur·euses de la HES-SO.
Développement de scripts Python qui s'appuient sur les Application Programming Interfaces (APIs) de ces services.

Pour le premier test: avec DataCite

Hes·so

MÉTHODES

Etapes principales

1. Identifier tous les jeux de données et codes/logiciels potentiels de l'institution
2. Filtrer et exclure les jeux de données et codes/logiciels liés à d'autres institutions
3. Regrouper les différentes versions d'un même jeu de données ou code/logiciel
4. Enrichir les jeux de données et codes/logiciels et évaluer leur FAIRness
5. Analyser les jeux de données et codes/logiciels collectés

1 Identifier tous les jeux de données et codes/logiciels potentiels de l'institution

Recherche par identifiants ROR

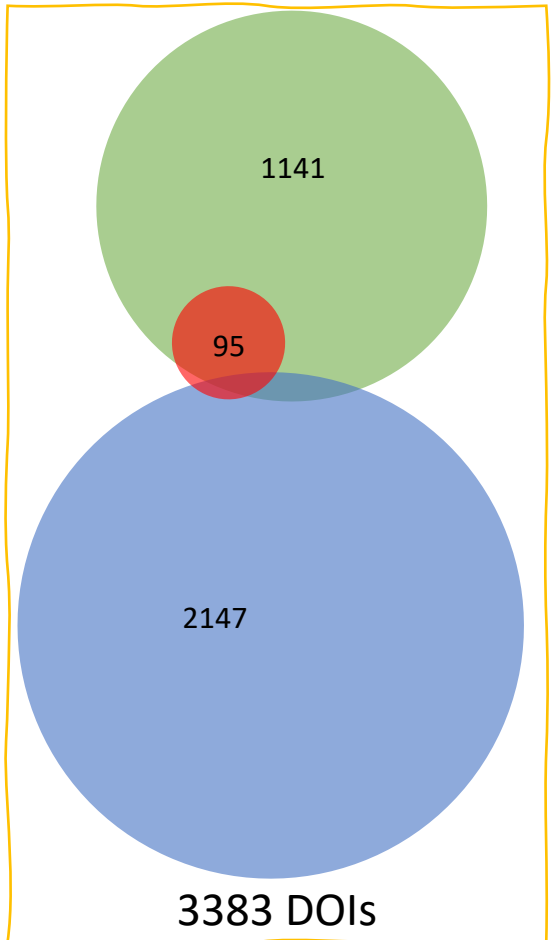
- \+ Excellente précision
- − Peu utilisé
- − DataCite n'étend pas automatiquement aux identifiants enfants (child organization).

Recherche par affiliations

- \+ Augmente le rappel
- − Moins bonne précision
- − Sensible à la casse
- − Beaucoup d'écritures possibles (institution/ école/institut, forme étendue/abrégée, français/anglais...)

Recherche par contenu

- \+ Augmente le rappel
- \+ N'est pas sensible à la casse
- − Mauvaise précision

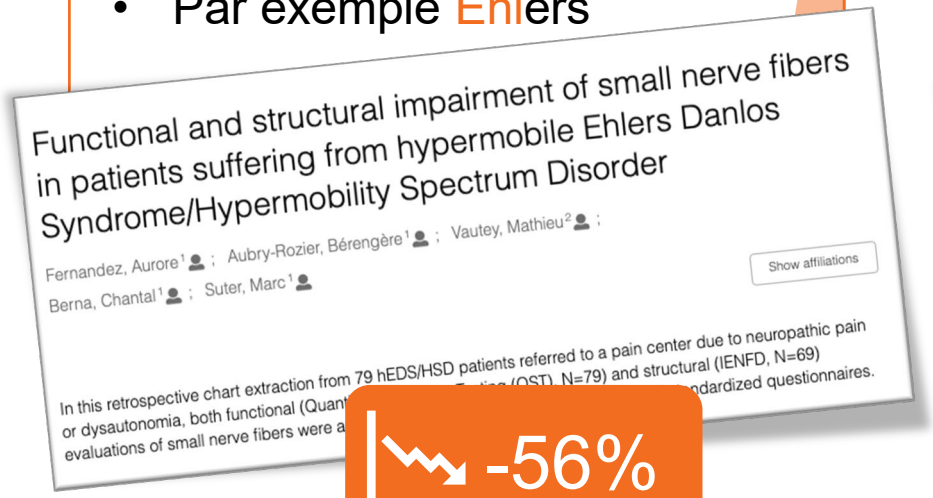


Répartition des DOIs

2 Filtrer et exclure les jeux de données et codes/logiciels liés à d'autres institutions

Exclusion selon le nom de l'affiliation

- Elimination des « matchs » inclus dans des mots: problème résultant de l'utilisation des wildcards.
- Par exemple Ehlers


Functional and structural impairment of small nerve fibers in patients suffering from hypermobile Ehlers Danlos Syndrome/Hypermobility Spectrum Disorder

Fernandez, Aurore[1] ; Aubry-Rozier, Bérengère[1] ; Vautey, Mathieu[2] ;
Berna, Chantal[1] ; Suter, Marc[1]

Show affiliations

In this retrospective chart extraction from 79 hEDS/HSD patients referred to a pain center due to neuropathic pain or dysautonomia, both functional (Quant (QST), N=79) and structural (IENFD, N=69) evaluations of small nerve fibers were a ndardized questionnaires.

-56%

Exclusion selon les noms des auteurs

- Recherche des auteurs dans ArODES: récupération des hautes écoles et domaines
- Exclusion des auteurs non retrouvés

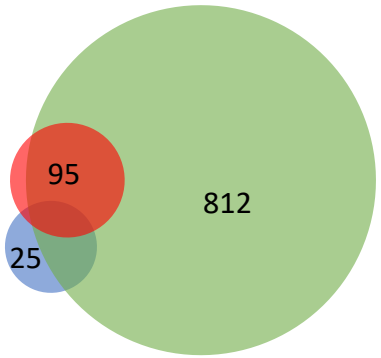

Laser Response in ECAL Crystals in CMS Detector

Bhargav Joshi[1] ; Roger Rusack[1]

Show affiliations

The dataset contains the Laser responses of the Lead-Tungstate crystals in the Electromagnetic Calorimeter (ECAL) of the CMS Experiment recorded during the Run 2 (2016-2018) of LHC running. The datasets consists of two tar folders: one corresponding to the "plus" side of the detector and one corresponding to the "minus" side. Each folder contains files in csv format, each file corresponding to the histories of all crystals in each "ieta" ring. The detailed description of the columns can be found under the section names "dataset" on Github pages at https://fair-umn.github.io/fair_ecal_monitoring.

-32%



932 DOIs

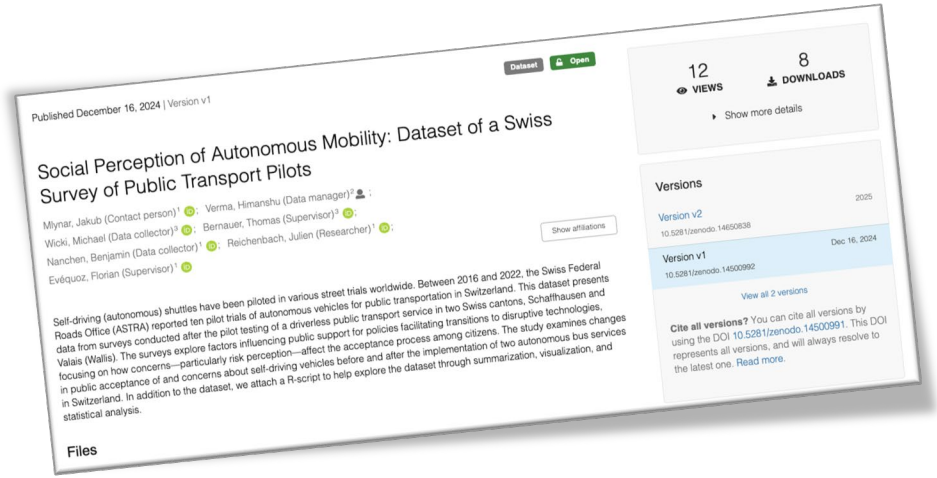
Répartition des DOIs

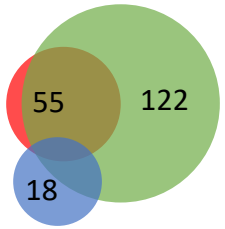
3 Regrouper les différentes versions d'un même jeu de données ou code/logiciel

Regroupement des versions

- Chaque jeu de données ou code/logiciel est rattaché à un DOI parent. Utilisation de ce DOI et fusion des versions.
- Cas particulier: 3 jeux de données déposés dans Harvard Dataverse sont composés de > 100 parties.

-76%





195 DOIs

Répartition des DOIs

4 Enrichir les jeux de données et codes/logiciels et évaluer leur FAIRness

Récupération de métadonnées supplémentaires

- Uniquement pour Zenodo
- DataCite ne retourne pas toutes les métadonnées ou certaines sont incomplètes/incorrectes
- Droits d'accès, types de fichiers, licences, nombres de téléchargements, etc.

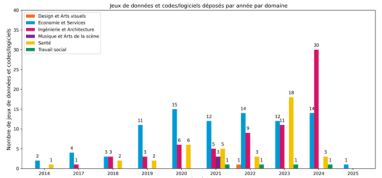
Evaluation de la FAIRness

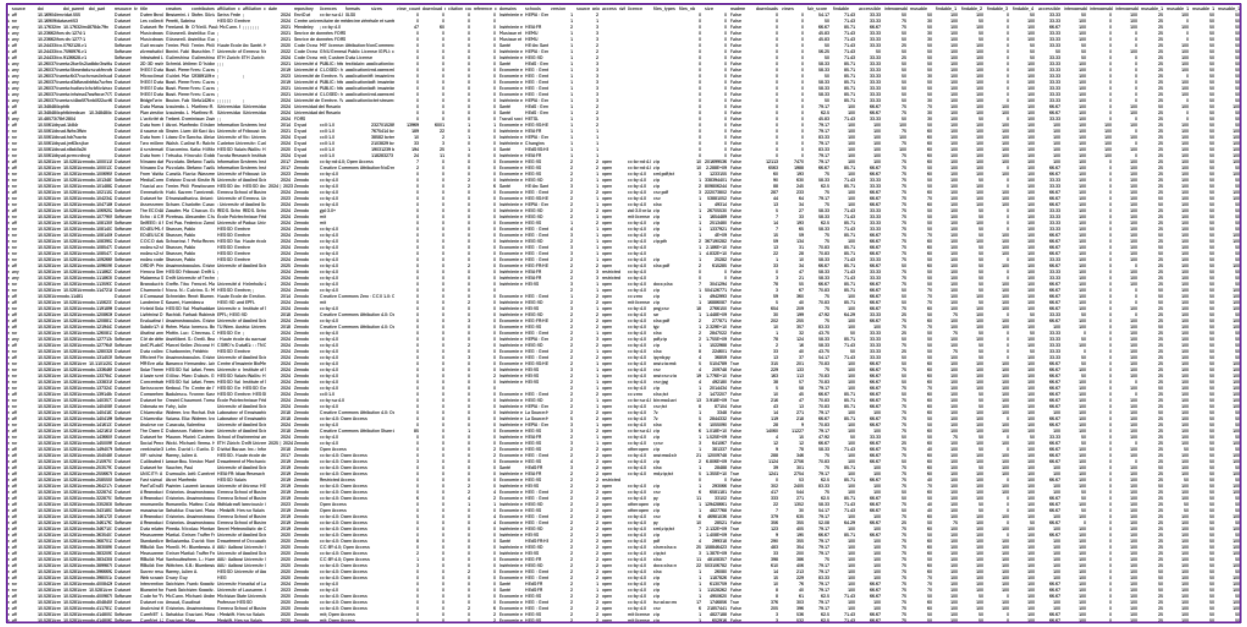
- Utilisation de F-UJI
- Basé sur la métrique « FAIRsFAIR Data Object Assessment »
- Divers scores récupérés: score général, score pour chacune des caractéristiques et sous-caractéristiques FAIR.

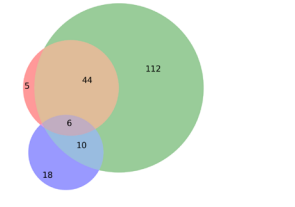
Anusuriya Devaraju, & Robert Huber. (2020). F-UJI - An Automated FAIR Data Assessment Tool. Zenodo. https://doi.org/10.5281/zenodo.6361400

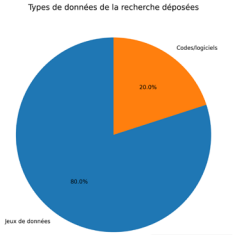
Hes·so

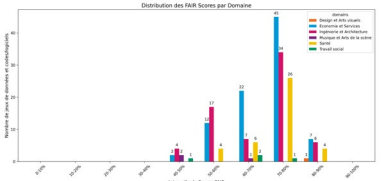
5 Analyser les jeux de données et codes/logiciels collectés

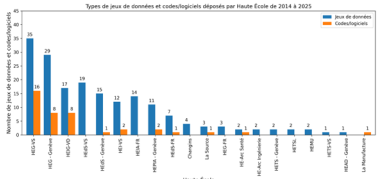


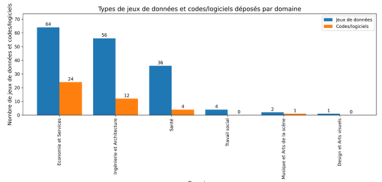


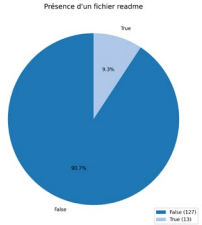


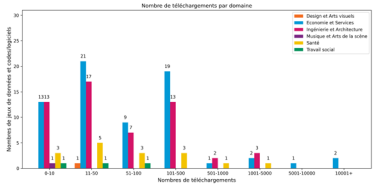


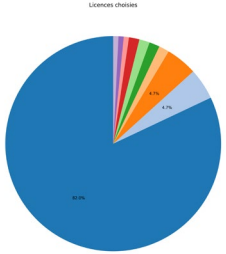












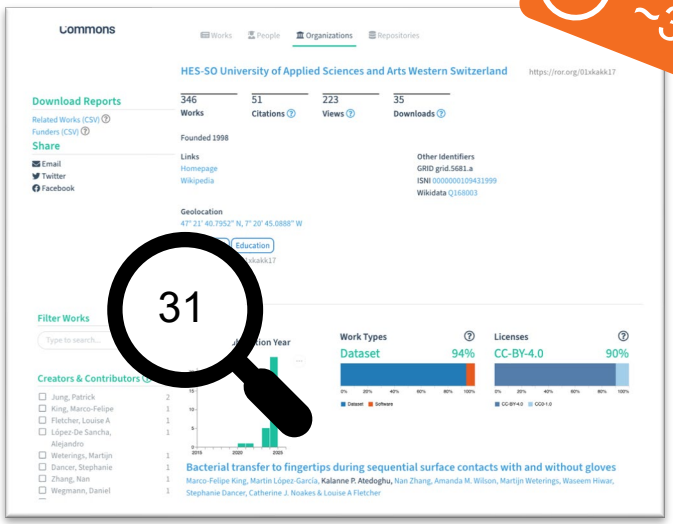
RÉSULTATS

En bref

Identification de **195** jeux de données et codes/logiciels distincts correspondant à **932** versions (DOI).

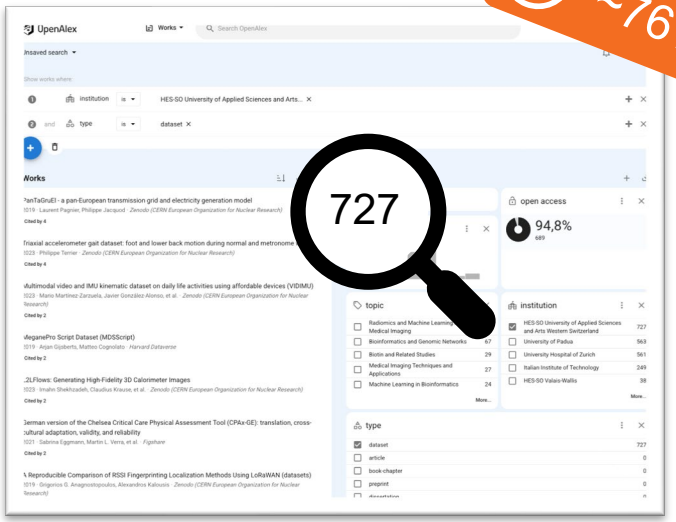
Comparaison DataCite vs. OpenAlex vs. Monitoring (Datacite étendu)

DataCite Commons



~3%

OpenAlex



~76%

Ce monitoring (DataCite étendu)

~98%

Hes·so

Comparaison OpenAlex vs. Monitoring (Datacite étendu)

Ce que ce monitoring (« Datacite étendu ») rate

9 datasets (19 DOIs):

- 4 datasets (8 DOIs) avec une affiliation de mauvaise qualité
- 3 datasets (8 DOIs) pas clairement rattaché à la HES-SO
- 1 dataset (2 DOIs) éliminé lors de l'étape 2: auteur absent dans ArODES
- 1 dataset (1 DOI) représentant un article

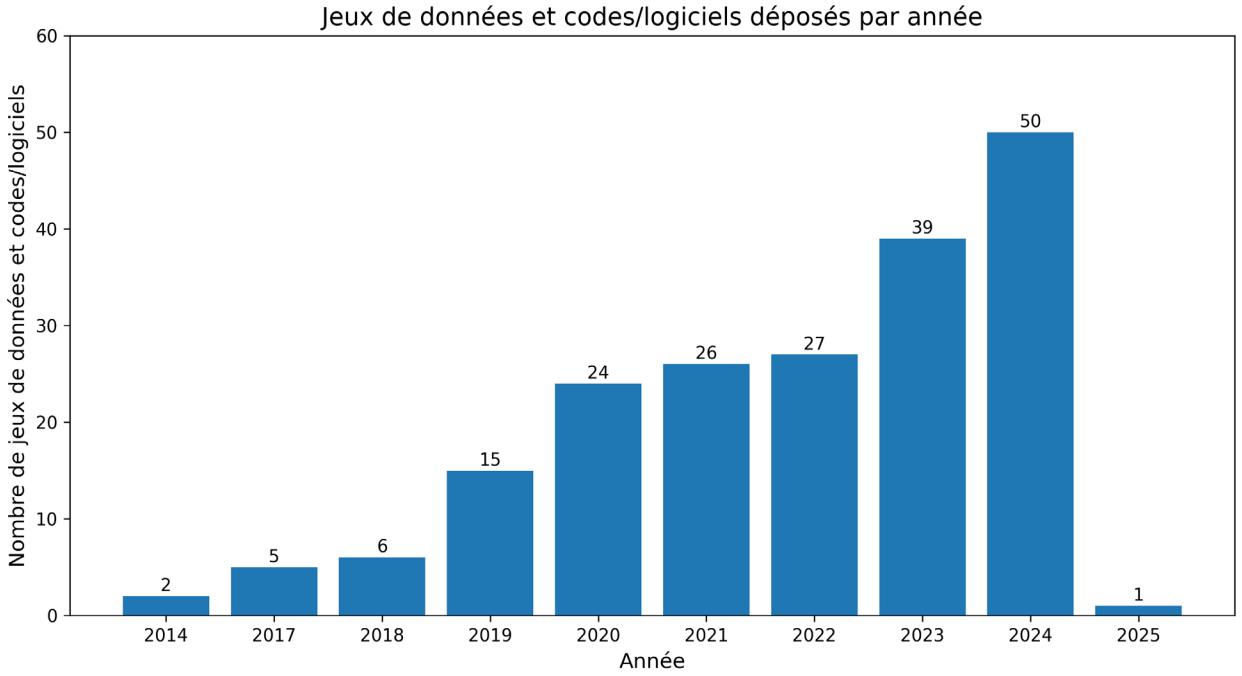
Ce que OpenAlex rate

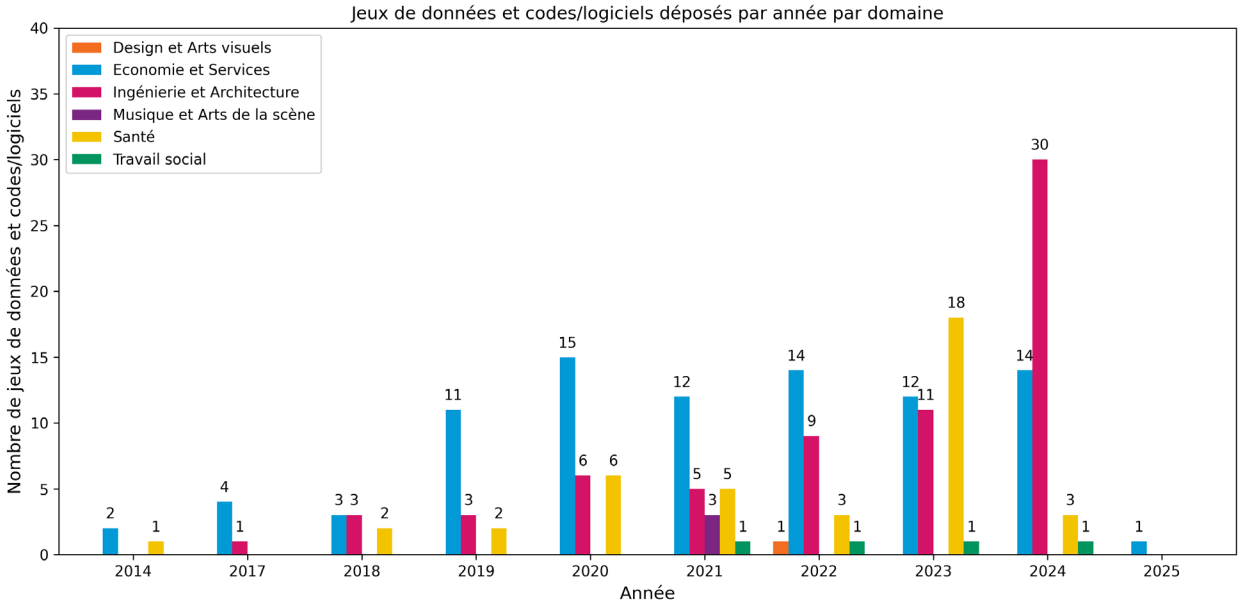
121 datasets/logiciels (220 DOIs):

- 39 codes/logiciels (84 DOIs) (codes et logiciels non représentés dans OpenAlex)
- 64 datasets (104 DOIs) non indexé par OpenAlex
- 18 datasets (32 DOIs) non rattaché à la HES-SO par OpenAlex

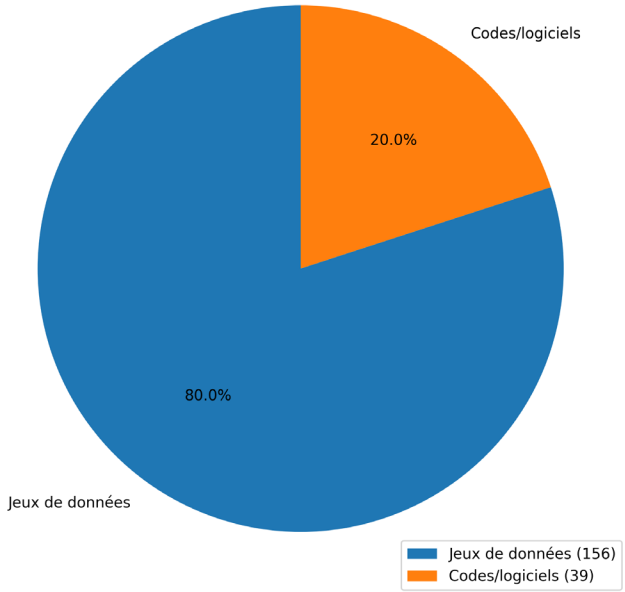
Hes·so

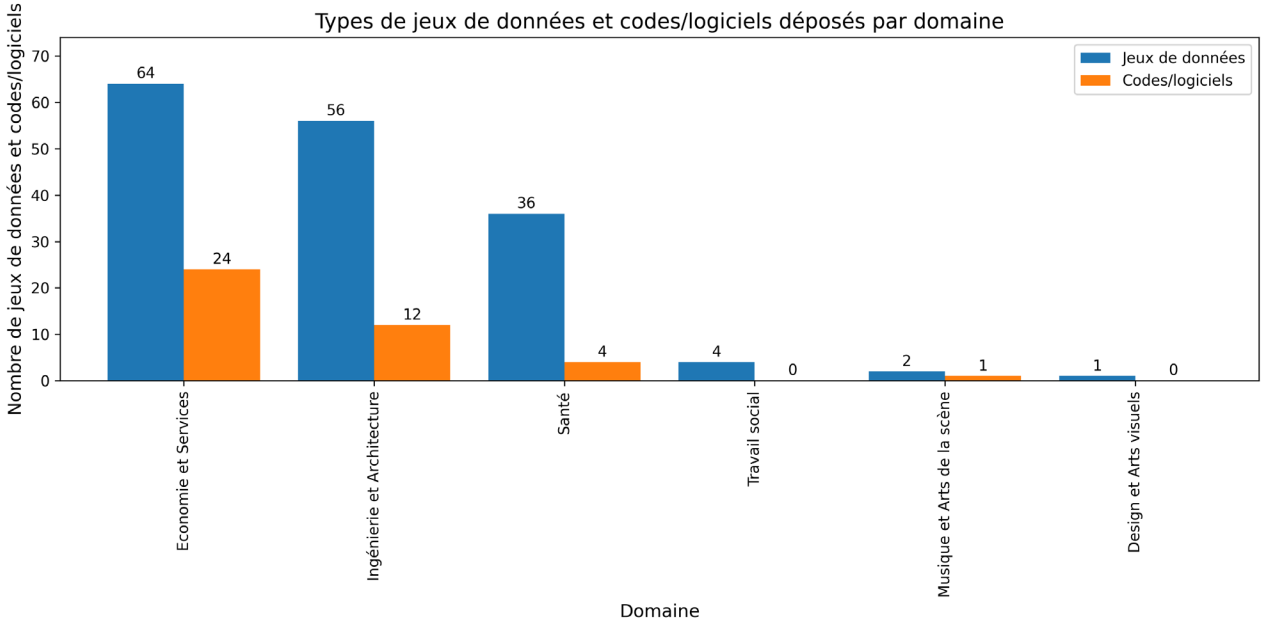
Jeux de données et codes/logiciels déposés par année



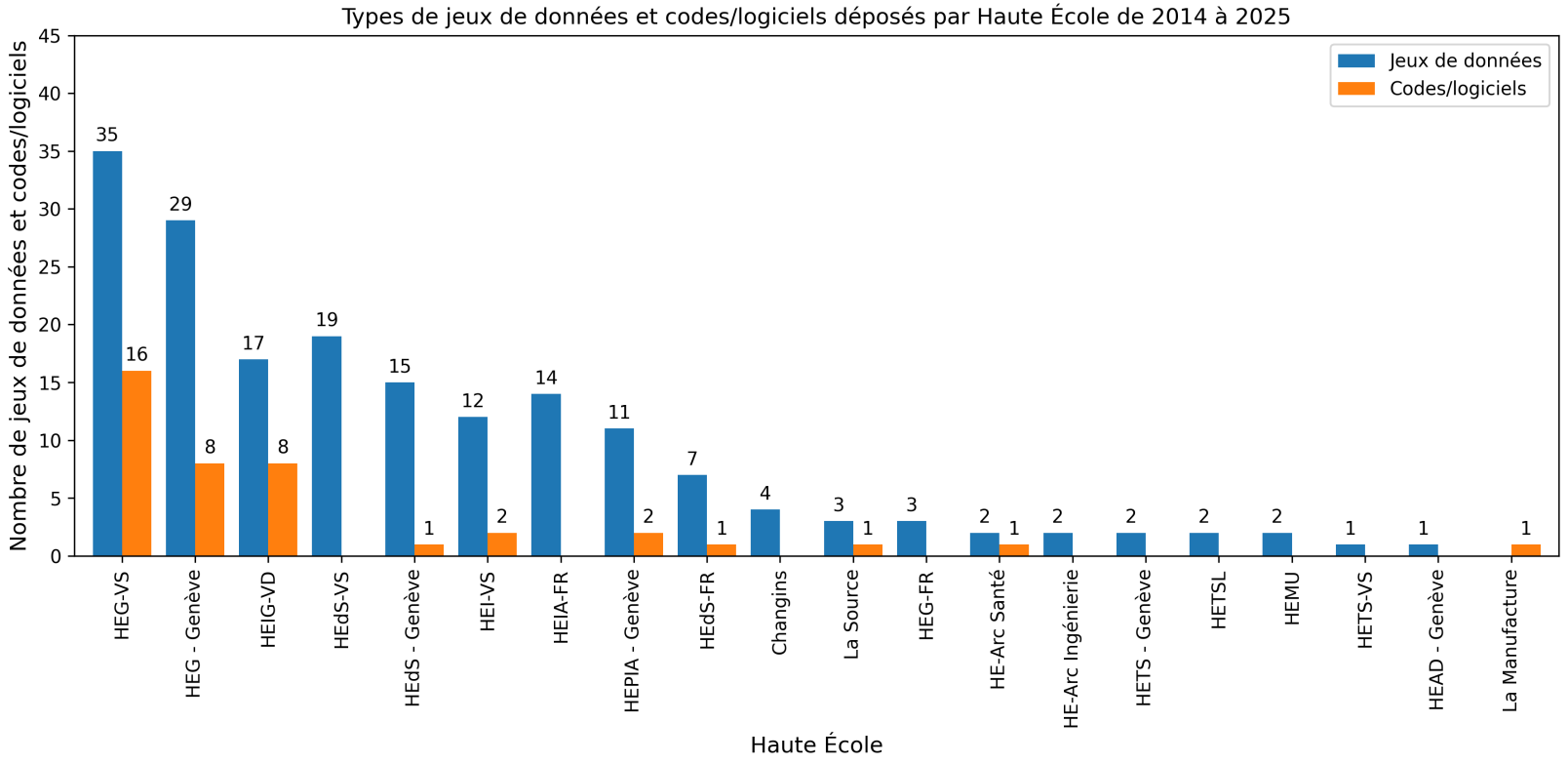


Jeux de données et codes/logiciels



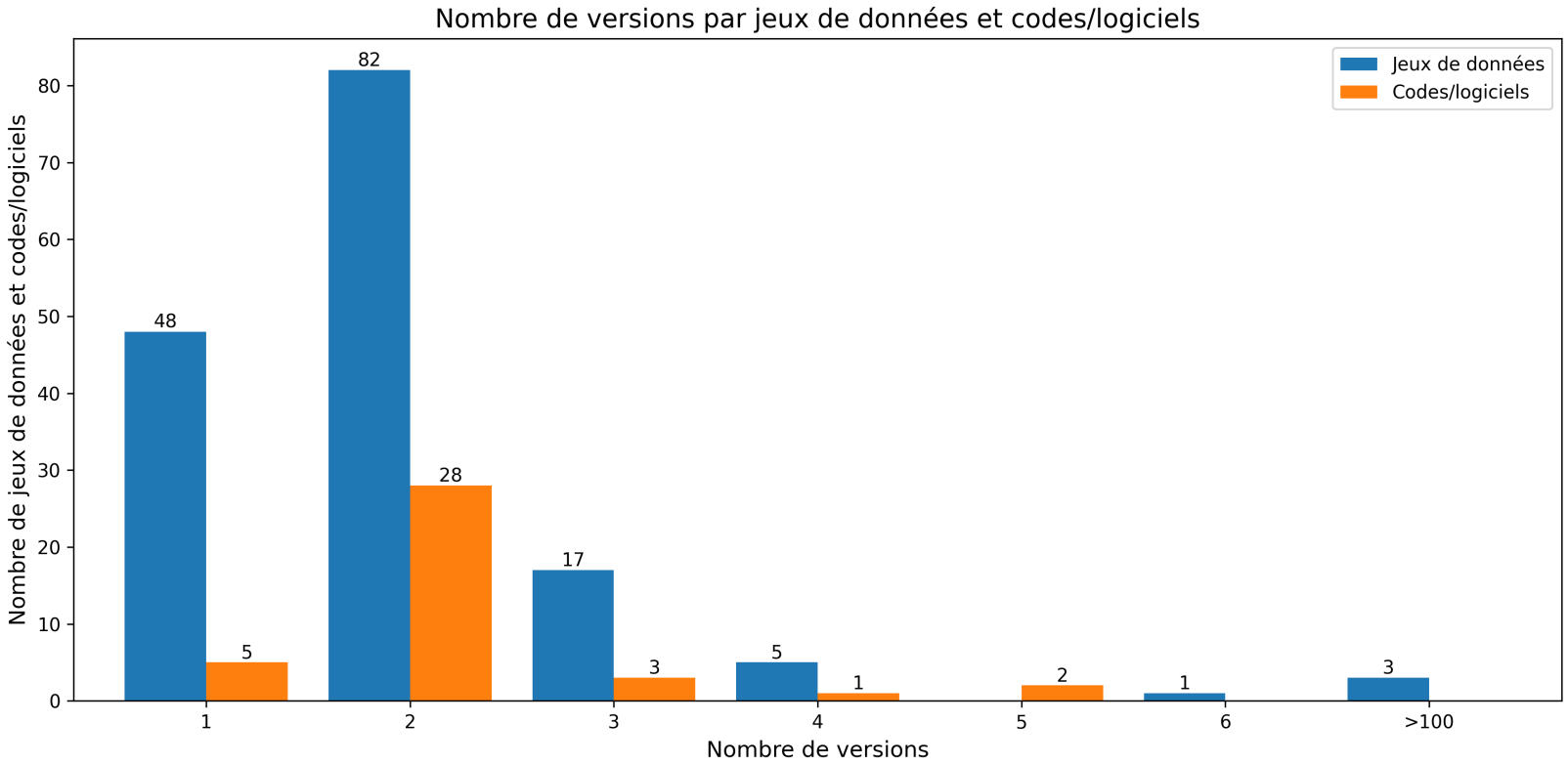


Jeux de données et codes/logiciels déposés par haute école de la HES-SO

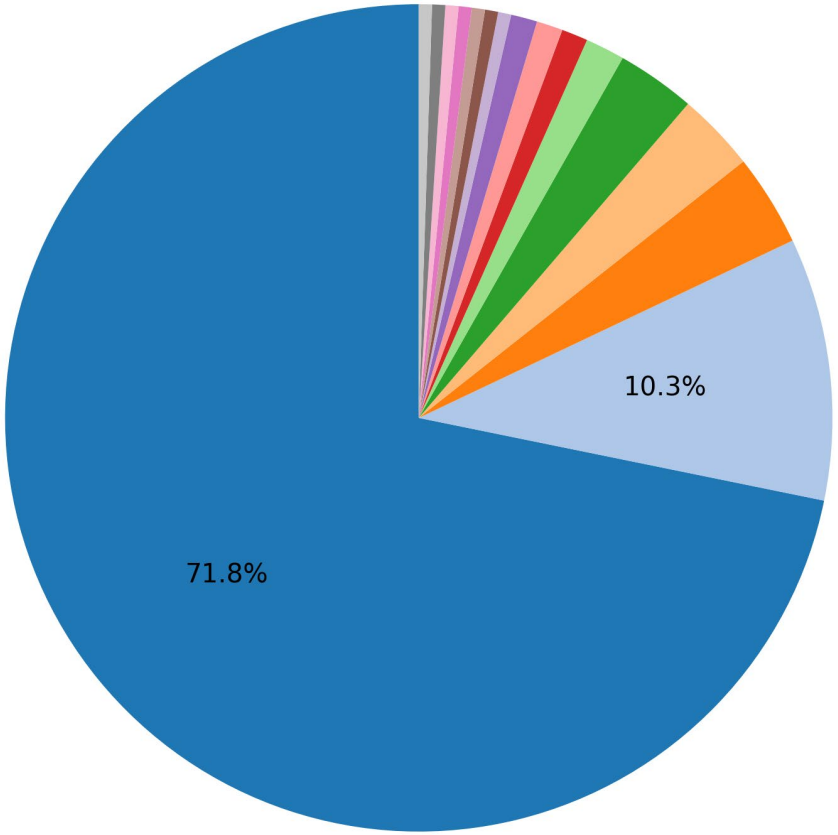


Note: aucun jeu de données ou code/logiciel n'a été trouvé pour les Hautes Ecoles non mentionnées dans ce diagramme.

Nombre de versions par datasets

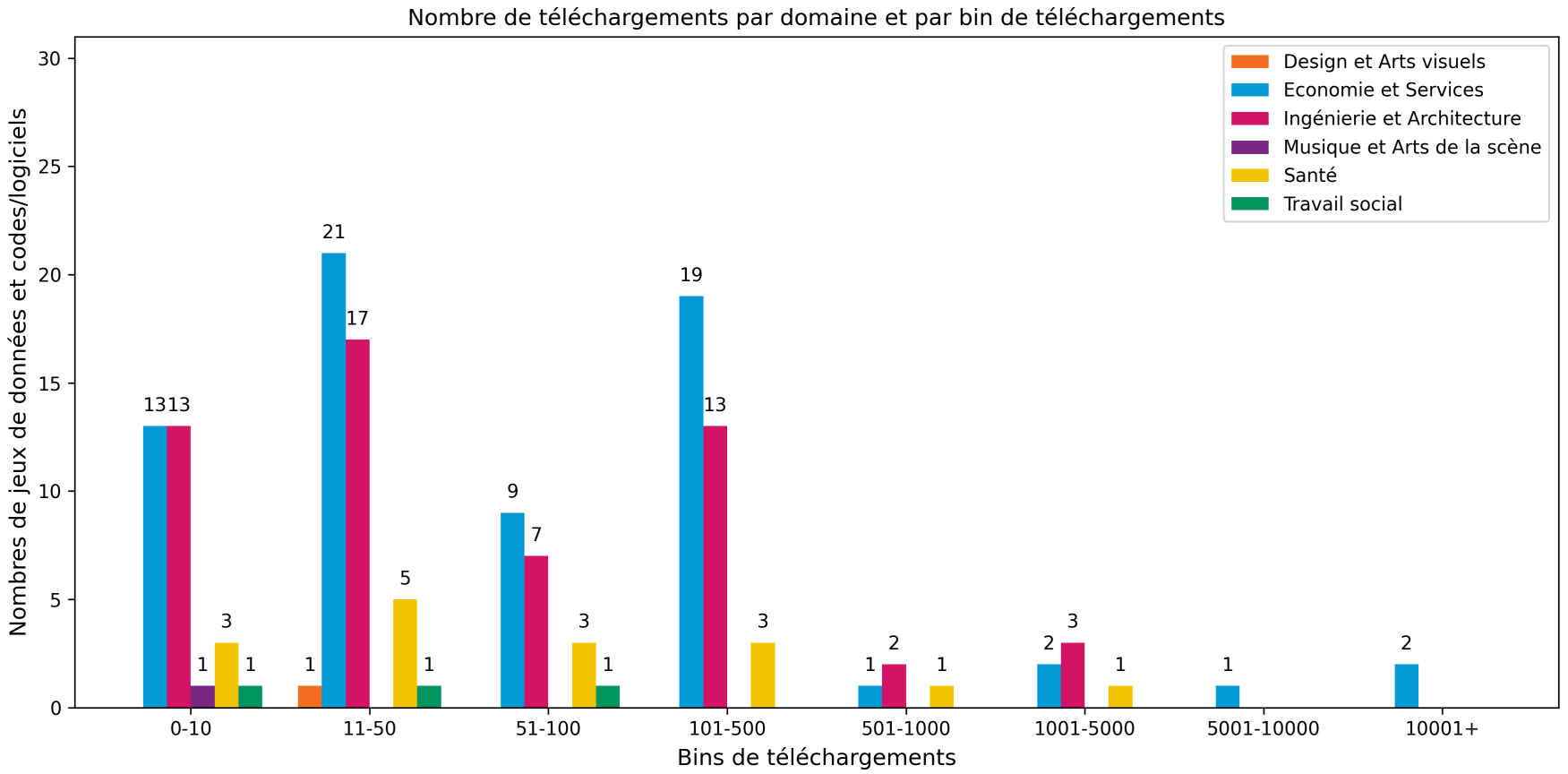


Dépôts choisis



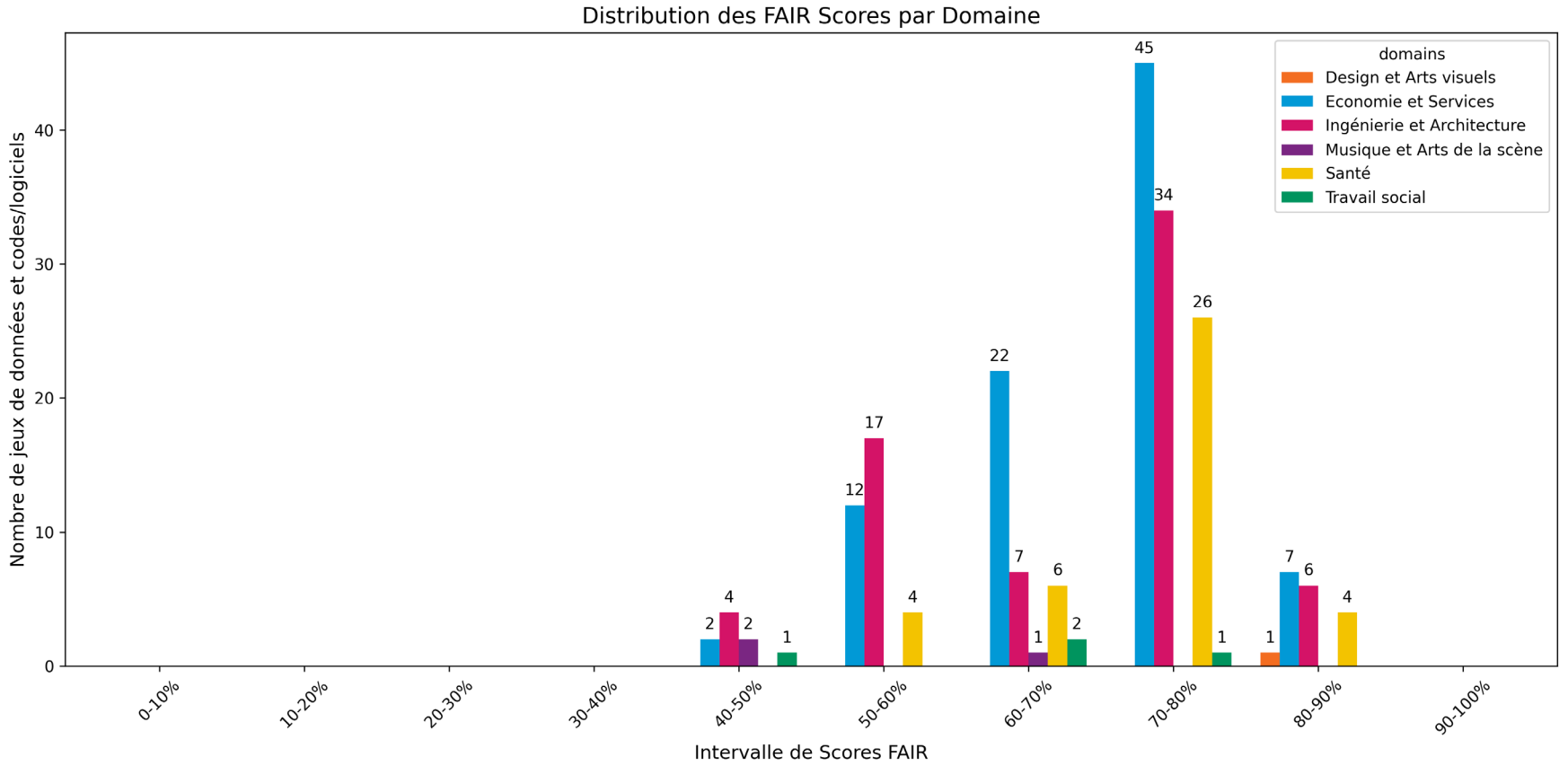
- Zenodo (140)
- figshare (20)
- Université de Genève, Yareta (7)
- Dryad (6)
- Harvard Dataverse (6)
- Code Ocean (3)
- Service de données FORS (2)
- Universidad del Rosario (2)
- Taylor & Francis (2)
- EnviDat (1)
- Centre universitaire de médecine générale et santé publique, Lausanne (1)
- Mendeley (1)
- FORS (1)
- University of Leeds (1)
- Borealis (1)
- Science Data Bank (1)

Nombre de téléchargements



Uniquement basé sur les dépôts dans Zenodo

FAIRness des jeux de données et codes/logiciels



CONCLUSION

Résumé des points clés

195 jeux de données et codes/logiciels ont pu être trouvés pour la HES-SO, dont environ la moitié a été déposés au cours des **2 dernières années**.

⇒ **Évolution positive des pratiques de partage des données**

⇒ Pratique qui a particulièrement évolué pour le domaine Ingénierie et Architecture

⇒ Pratique plutôt stable au cours des 6 dernières années pour le domaine Economie et Services

⇒ **Méthode provisoire** pour avoir un aperçu des tendances

Résumé des points clés

Limitations:

- Manque de standardisation des affiliations rend le monitoring compliqué
  - Nécessité de mieux communiquer sur les affiliations attendues pour les dépôts de jeux de données et logiciels, comme c'est le cas pour les publications.

- Les APIs de DataCite ne sont pas optimales pour la recherche par affiliations (sensibilité à la casse, wildcard)
  - Nécessité de connaissances en programmation pour contourner ces limitations

- OpenAlex est plus performant pour retrouver les jeux de données associés à une institution, mais à une couverture pour le moment plus faible que DataCite.
  - En particulier, pas de codes/logiciels

Disclaimer

La méthode présentée dans le cadre de cette présentation, fournit une estimation des pratiques de partage des données, mais elle ne garantit pas un recensement exhaustif. Les résultats obtenus dépendent des sources d’agrégation utilisées et des limites inhérentes aux outils d’indexation, pouvant entraîner une sous-représentation des pratiques de partage réelles.

Hes·so